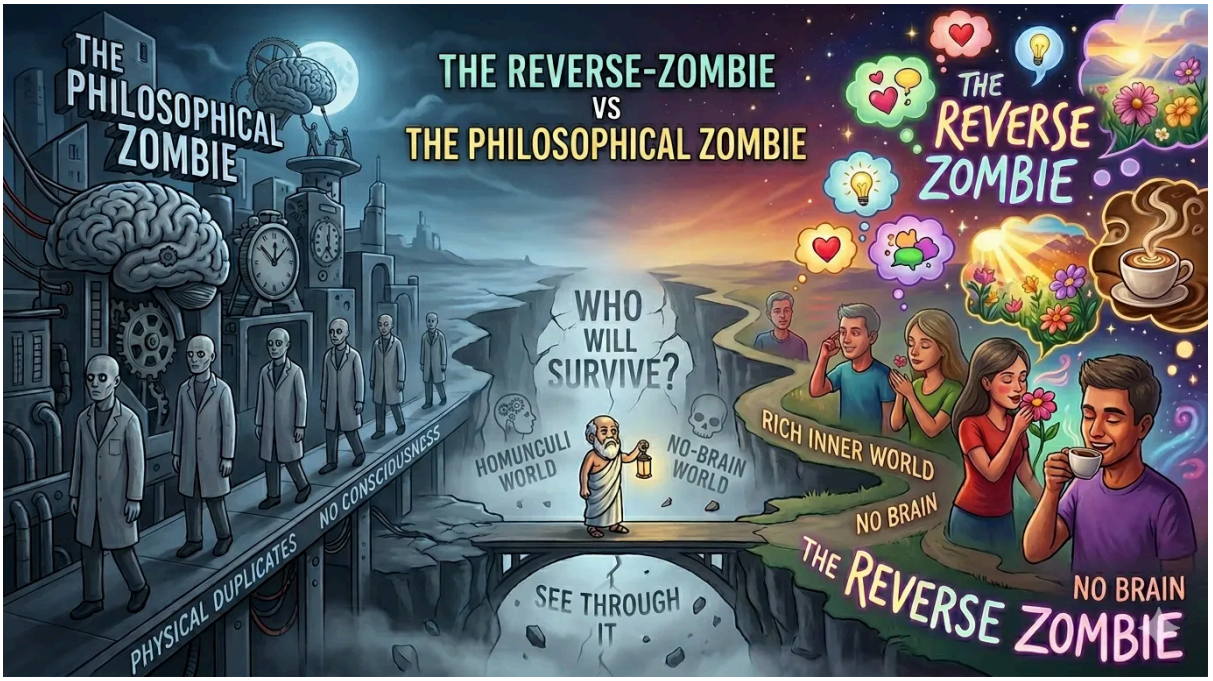
The Reverse-Zombie Vs The Philosophical Zombie: Who Will Survive?



"*Maybe there's Two-Brain World, Homunculi World, Pixie World, or even No-Brain World. None of these are logically impossible, but if logical conceivability means 'anything you can imagine' then it has no business doing philosophy.*"

This is part of a response I wrote to a comment left on my previous essay: '*How to Kill a Philosophical Zombie*'. When I wrote that essay I had one single mission: to prove beyond any doubt that the p-zombie argument had produced nothing of philosophical significance in thirty years and ought to be abandoned as soon as possible (and ideally removed from the syllabus that I have some responsibility for teaching).

I failed, obviously, not because the argument was too strong — almost the opposite — because it was barely even an argument. There's no clean headshot that will take it out. It is immune by design and encased in such heavy plot armour — the kind that implausibly keeps a protagonist alive in fiction — that there's no way through to that zombie brain.

This essay has a different purpose: to show the argument is just a mirage after all, with no substance. You can't dismantle a mirage. All you can do is see through it and keep walking.

Here's a summary of the argument itself: imagine a being physically identical to you in every way — same neurons, same brain structure, same behaviour — but with nobody home, no

inner experience. The lights are on, everything is running as it should, but there is no awareness, nothing it is like to be this creature. They are a philosophical zombie.

According to David Chalmers a p-zombie is logically conceivable. If it is logically conceivable, then it is metaphysically possible. If it is metaphysically possible, then consciousness cannot be purely physical — because here is a physical duplicate that lacks it entirely. Physicalism is false, and consciousness is something over and above the physical. The hard problem of consciousness is a genuine and perhaps permanent mystery.

I'm not the first person to say such a being is not actually conceivable. I think we tend to feel it straight away - the intuition that it is not actually conceivable. Then this gut feeling gets buried in arcane modal logic, and esoteric psychophysical laws. You have to be a specialist who has invested significant amounts of time studying and reflecting to even understand the nuances.

Like a lot of people, Socrates included, I think there is something wrong with that kind of philosophy. If something doesn't make sense unless it is padded in hundreds of words of specialist phrases, then the bare bones of it is probably not a good argument. In the context of p-zombies, let me explain why.

In thinking about the comments to my original essay I discovered another possible world. Reverse Zombie World is its provisional name. We are there, you and I, you reading these words and me writing them. The key difference is this: our reverse zombie doppelgangers still have the full richness of our inner world. They share our hopes and fears, mine is as sceptical of p-zombies as I am. They see red, smell flowers and taste coffee. But they have no brain inside their skull. It is hollow. No grey matter whatsoever. And here's the thing: they pass Chalmers' test of logical conceivability in just the same way as a p-zombie does. There is no logical contradiction. Does that mean that minds can float free of physical constraints? Does Reverse Zombie World say anything substantive about the nature of consciousness? Of course not. It is ridiculous.

Now imagine that two worlds to the left, in the hypothetical multiverse, there is another world. Same basic setup, only this time everyone has a little homunculi inside their skull, pulling levers, talking into microphones, operating machinery. Is Homunculi World logically

conceivable? How about Transcendental Pixie World next door, or even Spaghetti Brain World. Or Two Brain World, next to Three and Four Brain Worlds. None of these possible worlds contains a logical contradiction, which means they are all logically conceivable. Or maybe the phrase is being used to do something it cannot really do in all of these formulations, including the p-zombie.

A possible response is that the reverse zombie and associated worlds represent a more radical departure from the actual world than the standard p-zombie. P-Zombie World holds all physical facts constant and only removes consciousness — one variable changed, everything else identical. The reverse zombie removes the brain entirely, and the others complicate things at least as much.

But conceivability, by Chalmers' own definition, is not about how close a scenario is to the actual world. It is binary — either there is a logical contradiction or there isn't. The reverse zombie contains no logical contradiction. A zombie enthusiast cannot introduce a sliding scale of acceptable departure from actuality without introducing the very empirical constraints that the p-zombie cannot live with. You cannot invoke empirical constraints selectively — letting them block the reverse zombie while refusing to let them block the p-zombie.

The Reverse Zombie isn't a perfect mirror to the p-zombie of course — consciousness is known directly from the inside while brains are known structurally from the outside. Arguably this asymmetry justifies treating the two conceivability claims differently. But this is exactly the plot armour I am talking about. If we decide that consciousness occupies a special non-physical domain before we even run the conceivability test, the test is just a theatrical performance of our own prior commitments. At every additional level of complexity this fundamental circularity survives.

Once you accept that the conceivability test operates independently of empirical reality, you have no principled way to stop anything. There is a vast expanse of things we can imagine without logical contradiction that are impossible in the real world of physics and biology. The argument rests on a strange and groundless assumption: that logic beats physics and biology as the fundamental law of what can exist. It assumes that because we can separate these concepts in the 'dictionary' of our minds, nature is obliged to allow them to separate in reality.

This is *Modal Rationalism* at its most overconfident — the belief that our imaginative capacity is a perfect map of metaphysical possibility. If it were, then anything we can imagine, no matter how bizarre, has its own possible world, so long as there is no internal inconsistency. Would the concept of 'possible worlds' still mean very much in that context?

There is a misuse of the phrase logical conceivability here that is the result of a category error, one that often accompanies ontological claims reliant on modal logic. The reasoning goes something like this:

"*I have the concept of a mind. I also have the concept of a brain. When I reflect on whether the mind could be absent from the brain or vice versa I find that nothing in my concept of either contradicts this. I therefore conclude that the mind and the brain are distinct from one another in the actual world".*

Do you see the error? It is often glossed over, but that final step is invalid — it attempts to leap from the internal coherence of concepts onto the things that those concepts are placeholders for. If the idea of a brain still makes sense with no mind, that tells us nothing of the relationship between an actual mind and an actual brain. The zombie argument may well tell us something about our concepts of mind and brain. But it says nothing significant about whether consciousness and brain can *actually* exist independently of each other.

The same thing happens when Plantinga and Malcolm try to use modal logic to bring God into existence, as if He could be created from a riddle. It's almost a reverse ontological argument that has parallels with the reverse zombie argument under consideration. You define something first. You see if your definition holds, and then you assume that what is true of your definition is somehow true of the thing you defined. No concept can hold God, just as no concept could hold the mind. We have no idea whether a brain could exist without a mind. All we have is the overwhelming empirical evidence that it is so unlikely as to be unworthy of consideration.

Science will be unlikely to ever explain why consciousness feels the way it does. Even if it progresses so far as to tell us everything about consciousness, its language could still not describe anything about how it feels or why. That isn't a limitation of science. It is just expecting a tool to do something it was not designed for. Science describes structure and

mechanism. The felt quality of experience is a different kind of question. That doesn't make it a permanent mystery — it makes it the wrong question to ask of science. The hard problem isn't hard because science has failed. It's hard because the question is misdirected.

I see a bad argument wearing such fancy clothes that it manages to convince people that it is good. Chalmers never demonstrated why a p-zombie is conceivable, however intricate the arguments became, and that basic suspicion is usually there when people first hear about the argument. It's the automatic reaction that then gets buried in a thousand words.

That is why it survives — it is genuinely intimidating. Modal logic, Kripke, natural versus logical necessity, possible worlds semantics — the machinery is elaborate enough that most people assume they must be missing something to doubt it. The sophistication is doing defensive work. It makes the argument look more formidable than it is.

The zombie argument is extraordinarily productive, partly because Chalmers is genuinely brilliant at presenting it in its most compelling form and anticipating objections. Philosophical skill deployed in service of a bad argument is still a bad argument. But it is harder to see that it is bad.

It also speaks of a kind of intellectual bankruptcy at the heart of the rational philosophy of mind. In a flourishing discipline an argument like this would have caused a minor ripple at best. It wouldn’t have dominated the field since the end of the previous century. But where else can we go? Behaviourism didn’t work and functionalism acts as if ontology doesn’t matter for philosophy and so explains nothing philosophical. Token identity theory, eliminative materialism are harder to argue against, but also philosophically inert. They suggest there is no mystery and no significance to consciousness.

Chalmers has a strong intuition — one shared by many people — that consciousness is something special. That there is something it is like to be me that no physical description can fully capture. That awareness doesn't feel like the same kind of thing as atoms and molecules, so it cannot possibly be explained in purely physical terms. This is not an unreasonable intuition. Consciousness is remarkable. The fact that anything feels like anything at all is genuinely strange.

But intuition is not a philosophical argument. The p-zombie is the kind of philosophy that starts with a conclusion and reverse engineers its premises. The argument is a construct, a vehicle designed to carry an intuition. The conceivability test formalises that intuition — gives it the appearance of rigour — without actually doing any philosophical work. The conclusion was present at the beginning. The apparatus was built around it afterwards.

The philosophical zombie was never quite alive as an argument — just a body animated by specialist apparatus. The reverse zombie was always there, waiting to be pointed at. A mind without a brain is no less conceivable than a brain without a mind. If one establishes a possible world, so does the other. If neither does, the conceivability test has proved nothing.

‘How To Kill A Philosophical Zombie’ never earned its title because you can’t. There is no way because it doesn't exist. Chalmers has spun a mirage and hypnotised philosophy of mind for three decades. It's time to wake up.